DOCUMENT RESUME

ED 074 095 TM 002 458

AUTHOR Swineford, Frances
TITLE An Assessment of the Kuder-Richardson Formula (20) Reliability Estimate for Moderately Speeded Tests.
PUB DATE Feb 73
NOTE 10p.; Paper presented at NCME Annual Meeting, New Orleans, Louisiana, February 28, 1973

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Aptitude Tests; *Scoring Formulas; *Test Interpretation; *Test Reliability; *Timed Tests
IDENTIFIERS *Kuder Richardson Formula

ABSTRACT

Results obtained by the Kudar-Richardson formula (20) adapted for use with R-KW scoring are compared with three other reliability formulas. Based on parallel tests administered at the same sitting the KR (20) estimates are compared with alternate-form correlations and with odd-even correlations adjusted by the Spearman-Brown prophecy formula. Comparisons are also made between KR (20) estimates and alternate-form correlations obtained for tests administered after intervals of six to ten months. All the results justify the use of the Kuder-Richardson procedure with tests that show no more than moderate speededness. (Author)

AN ASSESSMENT OF THE KUDER-RICHARDSON FORMULA (20) RELIABILITY ESTIMATE FOR MODERATELY SPEEDED TESTS

Frances Swineford

Educational Testing Service

ABSTRACT

Results obtained by the Kuder-Richardson formula (20) adapted for use with R-KW scoring are compared with three other reliability formulas. Based on parallel tests administered at the same sitting the KR (20) estimates are compared with alternate-form correlations and with odd-even correlations adjusted by the Spearman-Brown prophecy formula. Comparisons are also made between KR (20) estimates and alternate-form correlations obtained for tests administered after intervals of six to ten months. All the results justify the use of the Kuder-Richardson procedure with tests that show no more than moderate speededness.

NCME Annual Meeting
New Orleans, La.
February 28, 1973

AN ASSESSMENT OF THE KUDER-RICHARDSON FORMULA (20) RELIABILITY ESTIMATE FOR MODERATELY SPEEDED TESTS

For some measurement specialists there continues to be doubt as to the appropriateness of the Kuder-Richardson formula (20) and its close relatives for estimating reliability unless all the examinees finish the test. This point of view raises a question of considerable importance because in large-scale testing programs it is frequently impractical to provide sufficient time to satisfy the slowest students. Consequently, assigned time limits are likely to represent a compromise between ideal power-test conditions and conditions that may introduce a moderate factor of speededness. The view that has been generally accepted at Educational Testing Service is that a test may be regarded as essentially unspeeded if at least 80 per cent of the examinees reach the last item and if virtually every one reaches three-quarters of the items. Some ETS tests do not quite meet both conditions. Nevertheless, the Kuder-Richardson formulas have been used with a high degree of confidence that they provide good estimates of test reliability. It is the purpose of this paper to present evidence that justifies that confidence. The Scholastic Aptitude Test happens to provide such evidence without any need for special testing. The conclusions to be drawn are properly restricted to test material similar to that of the SAT, although there is every reason to believe that generalizations can be made to other tests with similar speed characteristics.

KR (20) versus Alternate-Form, Same Administration

The analysis sample for each new form of the SAT is selected from the records of candidates who took one of the equating sections. Each equating section is a parallel form of one of the operational sections with respect to content, timing, and number of items. Listed in Table 1 are data for the two parallel sections, A and B, in thirty SAT forms. Sample sizes range from 370 to 2,000. From the per cents who reached three-quarters of the items, it is seen that our first condition for an unspeeded test is approximated for all the verbal sections and that the mathematical sections fail to meet it by about 1 to 4 per cent in general and by as much as 11.6 per cent in one instance. Instead of the per cent reaching the last item, our second condition for an unspeeded test, there has been recorded the number of items reached by less than 80 per cent of the group. These figures, too, suggest more speed in the mathematical scores than in the verbal scores.

Table 1

Comparison of Kuder-Richardson Formula (20) Estimates with Alternate-Form Correlations, Same Administration

Test	N	Per Cent Who Reached Three-quarters of Items		Number of Items Reached by Less Than 80 Per Cent of Group		KR (20) Reliability Estimate		r_{AB}
		A	B	A	B	A	B	
40-Item Verbal Sections								
1	900	99.7	98.9	2	2	.854	.878	.857
2	900	99.9	99.8	2	1	.826	.825	.809
3	900	99.4	99.4	1	0	.832	.850	.821
4	2,000	99.7	99.6	1	1	.828	.847	.833
5	900	99.4	99.6	3	1	.815	.844	.818
6	900	99.9	100.0	1	0	.825	827	.821
7	1,995	99.5	99.2	1	3	.849	.850	.839
8	845	99.8	99.9	0	1	.844	.869	.842
9	900	99.9	99.6	0	3	.851	.861	.848
10	900	98.6	99.4	2	1	.808	.863	.809
11	900	100.0	99.8	3	2	.815	.848	.815
12	495	99.4	100.0	2	2	.796	.848	.804
13	370	100.0	100.0	1	1	.828	.820	.833
14	370	100.0	100.0	2	2	.825	.855	.832
25-Item Mathematical Sections								
15	865	96.5	97.1	2	3	.825	.812	.818
16	1,885	96.2	88.4	2	6	.828	.791	.802
17	955	96.8	94.7	2	3	.781	.816	.789
18	900	98.6	96.9	2	2	.827	.807	.814
19	1,995	98.3	96.9	2	3	.850	.830	.832
20	900	96.3	98.1	2	3	.835	.833	.833
21	845	97.6	98.2	2	3	.823	.809	.798
22	900	96.1	95.9	3	3	.827	.813	.810
23	900	95.8	97.3	3	3	.801	.834	.808
24	900	95.8	98.6	4	1	.790	.844	.804
25	495	97.0	95.8	4	4	.803	.806	.815
26	495	92.5	98.2	5	2	.784	.784	.778
27	900	98.7	94.6	3	4	.793	.812	.815
28	370	98.9	97.6	1	1	.814	.833	.797
29	370	93.2	96.8	4	2	.780	.800	.806
30	370	92.7	95.1	4	2	.809	.789	.813

Inasmuch as all the tests involved in this study have been scored by the formula, Score = R-W/4 , Dressel's adaptation of the Kuder-Richardson formula (20), which renders it appropriate for use with such scoring, has been used (2). In no instance did the two sections, A and B, appear consecutively in the test. Seven of the verbal pairs were separated by additional verbal material; the other seven, by both verbal and mathematical sections. On the other hand, the two mathematical sections were all separated by additional mathematical material.

With these considerations in mind it is interesting to note that the alternate-form reliability, r_{AB} , is not always the lowest estimate, as might reasonably be expected, but, rather, it lies between the two K-R estimates in thirteen tests, equals one of them in three, and is higher than either in five more. For the verbal sections the mean value of r_{AB} is .827 and the mean KR (20) value is .839, whereas means for the mathematical sections are .808 for r_{AB} and .812 for KR (20).

As for the effect of speededness, there is no evidence to support the contention that the K-R estimate is inflated by the degree of speededness encountered in these tests. Quite the other way: for this particular set of thirty-two 25-item mathematical sections of Table 1, there is a positive correlation of .44 between the KR (20) reliability estimate and the per cent of the group who reached three-quarters of the items, and there is a negative correlation of .46 between the KR (20) estimate and the number of items reached by less than 80 per cent of the group. Thus the speededness shown for these 32 tests tends to be accompanied by slightly lower Kuder-Richardson estimates rather than higher values.

KR (20) versus Odd-Even, Same Administration

The data of Tables 2, 3, 4, and 5 are based on a single form of the SAT. Score A is a 40-item operational section and Scores C and D are 40-item equating sections that parallel A. Similarly, Scores B, E, and F are parallel 25-item sections. The four samples, of over 1,200 cases each, are mutually exclusive. The new data provided in these tables are intercorrelations, means, and standard deviations for scores on the odd-numbered and even-numbered items. In each table the "Total" rows contain in the last five columns the alternate-form reliability, the KR (20) estimates, the odd-even correlations stepped up by the Spearman-Brown prophecy formula, and, in the last column, the estimate that employs the odd and even variances and covariance. This last formula, attributed by Kelley (3) to

Table 2

Comparison of Kuder-Richardson Formula (20) Estimates with Estimates Based on Scores on Odd and Even Items

Sample 1. 40-Item Sections, 1,295 Cases

Score	Mean	S.D.	Intercorrelations						Reliability Estimate		
			A Odd	A Even	C Odd	C Even	A Total	C Total	(a)	(b)	(c)
A Odd	6.39	4.40		.735	.735	.754	.936	.797	.745		
A Even	6.26	4.08	.735		.704	.717	.923	.758	.696		
C Odd	7.28	4.38	.735	.704		.747	.772	.931	.733		
C Even	6.54	4.53	.754	.717	.747		.790	.934	.758		
A Total ...	12.55	7.91	.936	.923	.772	.790		.835	.840	.847	.846
C Total ...	13.68	8.31	.797	.758	.931	.934	.835		.854	.855	.855
Per cent who reached three-quarters of items							99.3	99.4			
Number of items reached by less than 80 per cent of group							3	1			

(a) KR (20) . (b) $\frac{2r_{OE}}{1 + r_{OE}}$. (c) $\frac{4C_{OE}}{C_{OO} + C_{EE} + 2C_{OE}}$.

Table 3

Comparison of Kuder-Richardson Formula (20) Estimates with Estimates Based on Scores on Odd and Even Items

Sample 2. 40-Item Sections, 1,270 Cases

Score	Mean	S.D.	Intercorrelations						Reliability Estimate		
			A Odd	A Even	D Odd	D Even	A Total	D Total	(a)	(b)	(c)
A Odd	6.42	4.53		.737	.748	.737	.940	.797	.761		
A Even	6.33	4.01	.737		.697	.695	.920	.747	.684		
D Odd	7.61	4.49	.748	.697		.742	.778	.933	.754		
D Even	7.04	4.42	.737	.695	.742		.771	.930	.745		
A Total ...	12.62	7.92	.940	.920	.778	.771		.831	.841	.849	.845
D Total ...	14.53	8.31	.797	.747	.933	.930	.831		.856	.852	.852
Per cent who reached three-quarters of items							99.0	99.8			
Number of items reached by less than 80 per cent of group							3	2			

See footnote to Table 2 for (a), (b), (c) references.

Table 4

Comparison of Kuder-Richardson Formula (20) Estimates with Estimates Based on Scores on Odd and Even Items

Sample 3. 25-Item Sections, 1,295 Cases

Score	Mean	S.D.	Intercorrelations						Reliability Estimate		
			B Odd	B Even	E Odd	E Even	B Total	E Total	(a)	(b)	(c)
B Odd	4.70	3.08		.693	.708	.706	.924	.770	.700		
B Even	4.57	2.82	.693		.695	.676	.908	.751	.680		
E Odd	4.96	3.17	.708	.695		.685	.763	.925	.701		
E Even	4.71	2.78	.706	.676	.685		.750	.903	.649		
B Total	9.14	5.38	.924	.908	.763	.750		.827	.814	.819	.817
E Total	9.53	5.45	.770	.751	.925	.903	.827		.807	.813	.809
Per cent who reached three-quarters of items							89.3	95.0			
Number of items reached by less than 80 per cent of group							5	3			

See footnote to Table 2 for (a), (b), (c) references.

Table 5

Comparison of Kuder-Richardson Formula (20) Estimates with Estimates Based on Scores on Odd and Even Items

Sample 4. 25-Item Sections, 1,275 Cases

Score	Mean	S.D.	Intercorrelations						Reliability Estimate		
			B Odd	B Even	F Odd	F Even	B Total	F Total	(a)	(b)	(c)
B Odd	4.87	3.21		.735	.705	.704	.936	.768	.731		
B Even	4.64	2.92	.735		.706	.669	.919	.750	.715		
F Odd	5.05	3.10	.705	.706		.686	.757	.917	.706		
F Even	5.07	3.03	.704	.669	.686		.739	.912	.695		
B Total	9.35	5.67	.936	.919	.757	.739		.815	.838	.847	.845
F Total	9.98	5.59	.768	.750	.917	.912	.815		.819	.814	.813
Per cent who reached three-quarters of items							86.8	94.7			
Number of items reached by less than 80 per cent of group							5	3			

See footnote to Table 2 for (a), (b), (c) references.

Flanagan, is equivalent to KR (20) with the item replaced by the half test as the basic unit. As Kelley notes, "The difference between this formula and [the Spearman-Brown formula] is trifling ...." In six of the eight possible comparisons of KR (20) with odd-even methods the latter provide slightly higher estimates than the Kuder-Richardson formula. Comparisons of the KR (20) estimates for the odd and even scores with the intercorrelations among those scores add nothing new to the findings but serve simply to support the evidence already provided in the data of Table 1.

KR (20) versus Alternate-Form, Different Administrations

Finally, data have been assembled from ETS files to compare the Kuder-Richardson reliability estimate that is provided for each new form of the SAT with alternate-form correlations obtained for candidates who, for various reasons, take a second form of the test after a period of six to ten months. Most of these candidates are juniors at the time of their first testing and seniors at the second. One example of this kind is given in Table 6, and another, for a different year, in Table 7. Together, these tables involve ten different forms of the test. Each verbal score and each mathematical score is based on two separately timed parts. The KR (20) reliability and its associated standard error of measurement are computed for each part and then combined by the expression, 1 - (error variance)/(total variance), where the error variance is the sum of the two variance errors of measurement and the total variance is the variance of the total score, to get the total-test reliability. The results of this expression are the values in the first five rows of each table. The verbal-score reliabilities for the ten forms range from .914 to .928, and the mathematical-score reliabilities, from .885 to .918.

The numerous factors that can affect correlations between measures separated by a substantial time interval are generally known and will not be covered here. To any one who desires a discussion of this subject the treatment by R. L. Thorndike in Lindquist (5) is recommended. Since the repeater groups of Tables 6 and 7 tend to be slightly less variable than the analysis samples, correlations based on the repeater groups may be expected to be correspondingly lower. In view of all the reasons one can find for expecting a drop in the alternate-form correlations, the data of Table 6 and 7 are particularly noteworthy. The mean verbal KR (20) reliability obtained for the ten analysis samples represented in the two tables is .92. Even the lowest

Table 6

Comparison of Kuder-Richardson Formula (20) Estimates with Alternate-Form Correlations, Different Administrations

First Example

Administration	Number of Cases	Verbal		Mathematical		r_{VM}	r_{VV}	r_{MM}
		Mean	S.D.	Mean	S.D.			
Analysis samples:								
March	900	461	109	487	113	.698	.921	.905
May	955	453	112	484	108	.685	.921	.885
November	2,000	458	106	495	108	.678	.916	.902
December	2,500	452	111	487	115	.682	.915	.905
January	900	434	114	466	114	.687	.928	.918
Repeater groups:								
March	45,843	481	103	511	105	.65-.68	.90	.88
November		491	105	528	106			
March	21,669	479	104	509	107	.65-.65	.90	.88
December		499	110	532	108			
March	3,262	458	108	490	111	.65-.68	.91	.89
January		474	112	504	113			
May	125,975	462	102	490	100	.63-.65	.89	.86
November		468	100	507	101			
May	66,421	459	104	490	104	.62-.65	.89	.87
December		477	105	514	106			
May	11,531	448	110	480	111	.67-.68	.90	.88
January		460	109	496	111			

Table 7

Comparison of Kuder-Richardson Formula (20) Estimates
with Alternate-Form Correlations, Different Administrations

Second Example

Administration	Number of Cases	Verbal		Mathematical		r_{VM}	r_{VV}	r_{MM}
		Mean	S.D.	Mean	S.D.			
Analysis samples:								
March	865	446	111	476	113	.708	.928	.911
April	900	455	108	481	111	.692	.914	.907
November	2,345	458	108	486	113	.683	.918	.907
December	2,000	448	108	481	110	.659	.915	.903
January	935	431	111	475	113	.681	.925	.916
Repeater groups:								
March	39,977	470	102	503	104	.66-.67	.90	.88
November		481	106	512	109			
March	18,669	472	104	504	108	.66-.68	.90	.88
December		493	108	524	109			
March	2,707	445	107	481	111	.69-.72	.91	.89
January		456	108	498	109			
April	117,975	463	101	493	104	.65-.65	.89	.88
November		469	103	501	107			
April	56,405	457	102	488	107	.64-.66	.89	.88
December		471	104	506	106			
April	9,906	443	108	472	111	.67-.71	.90	.89
January		447	107	491	107			

verbal alternate-form coefficient is as high as .89, and the mean of the twelve such coefficients is .90. Similarly, the mean mathematical KR (20) estimate is about .91, whereas the mean alternate-form coefficient is .88. These findings are perhaps the most compelling of all to justify confidence in the use of the Kuder-Richardson procedure with tests whose speed characteristics do not greatly differ from those of the SAT.

Selected References

1. Cronbach, Lee J. "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, XVI (September, 1951), 297-334

2. Dressel, Paul L. "Some Remarks on the Kuder-Richardson Reliability Coefficient," Psychometrika, V (December, 1940), 305-10.

3. Kelly, Truman L. "The Reliability Coefficient," Psychometrika, VII (June, 1942), 75-83.

4. Kuder, G. F., and Richardson, M. W. "The Theory of the Estimation of Test Reliability," Psychometrika, II (September, 1937), 151-60.

5. Lindquist, E. F. (editor). Educational Measurement, Chapter 15. Washington: American Council on Education, 1951.